On Document Classification with Self-Organising Maps

Jyri Saarikoski¹, Kalervo Järvelin², Jorma Laurikkala¹, and Martti Juhola¹

¹ Department of Computer Sciences, 33014 University of Tampere, Finland
² Department of Information Studies, 33014 University of Tampere, Finland

Abstract. This research deals with the use of self-organising maps for the classification of text documents. The aim was to classify documents to separate classes according to their topics. We therefore constructed self-organising maps that were effective for this task and tested them with German newspaper documents. We compared the results gained to those of k nearest neighbour searching and k-means clustering. For five and ten classes, the self-organising maps were better yielding as high average classification accuracies as 88-89%, whereas nearest neighbour searching gave 74-83% and k-means clustering 72-79% as their highest accuracies.

1 Introduction

The growth of digital documents and information stored as text in the Internet has been rapid in the recent years. Searching and grouping such documents in various ways have become an important and challenging function. A myriad of documents are daily accessed in the Internet to find interesting and applicable information. Distinguishing in some way interesting documents from the uninteresting ones is, even if a self-evident goal, crucial. For this purpose, computational methods are of paramount importance. We are interested in researching the classification of text documents, also those written in languages other than English.

There are known methods for constructing groups, clusters or models of documents, see for instance [4], [12] and [13]. These machine learning methods have included k nearest neighbour searching, probabilistic methods such as Naïve Bayesian classifiers [5] and evolutionary learning with genetic algorithms [13]. The methods were of the supervised category. We investigated the use of unsupervised Kohonen self-organising maps [8] that seemed to be seldom used in this field. They have been, however, applied to constructing visual maps of text document clusters, in which documents were clustered based on the features they contain. WEBSOM [7] [9] was employed to organize large document collections, but it did not include document classification in the sense to be compared with the current research. Chowdhury and Saha [2] classified 400, 500 and 600 sport articles, whereas Guerro-Bote et al. [6] employed 202 documents from a bibliographic database. Moya-Anegón et al. [10] made domain analysis of documents with self-organising maps, clustering and multidimensional scaling. Instead of
document clustering, we were interested in investigating how accurately and reliably self-organising maps are able to classify documents. Therefore, we constructed self-organising maps on document sets belonging to different known classes and used them to classify new documents. We employed ten-fold crossvalidation runs on our test document collection to assess classification accuracy in the document collection. We also performed comparable tests with $k$ nearest neighbour searching and $k$-means clustering which employ supervised learning to find a baseline level for the classification of the document data used. In principle, the use of self-organising maps is reasonable, because outside laboratory tests there is not necessarily a reliably classified learning set available.

In the present research, we extend our previous research of using self-organising maps for information retrieval in the same German document collection as in [11]. In the prior work, we studied retrieval from the document collection, the topics of which were associated with some of its documents, and we used both relevant and non-relevant documents in the document sample extracted from the collection. In the present work, our interest was in the classification, in other words separation between document classes. We therefore used only documents relevant to the classes examined.

2 The Data and Its Preprocessing

We used a German document set which was taken from an original collection of 294809 documents [1] from CLEF 2003 of the years 1994 and 1995 (http://www.clef-campaign.org/). The articles were from newspapers such as Frankfurter Allgemeine and Der Spiegel. There were 60 test topics associated with the collection. In every topic there was a relatively small subset of relevant documents. Relevant topics were included in our tests. At first, 20 topics were taken from the 60 topics otherwise randomly. From those 20 selected the smallest classes (topics) were still left out which included 6-25 relevant documents in the collection. Such small document classes would not have been quite reasonable for 10-fold crossvalidation tests, because their average numbers of test documents in test sets would only have been from 0.6 to 2.5, which might have resulted in considerable random influence. Thus, we attained 10 topics (classes) and 425 relevant documents (observations) so that the numbers of the relevant documents of the topics were 27, 28, 29, 29, 34, 39, 44, 53, 55 and 87.

The concept of relevance means here that the association of the documents to the topics had been manually ensured in advance by independent evaluators who had nothing to do with the present research.

To transform pertinent document data into the input variable form for a self-organising map, some preprocessing was required. At first, the German stemmer called SNOWBALL was run to detect word stems like ‘gegenteil’ from ‘Gegenteil’ or ‘gegenteilig’ from all documents and topics chosen. In addition, a list of 1320 German stopwords was used to sieve semantically useless words from them. Stopwords are prepositions like ‘ab’, articles like ‘ein’ and ‘eine’ or pronouns like ‘alle’, adverbs or other uninteresting “small words”, which are mostly uninflected words. They were removed from the documents and topic texts. Thereafter, short words, shorter than four letters, were removed, because they are typically, after stemming, as word